Naif Arab University for Security Sciences
Journal of Information Security and Cybercrimes Research
مجلة بحوث أمن المعلومات والجرائم السيبرانية
https://journals.nauss.edu.sa/index.php/JISCR
JISCR

AI Hallucinations in Information Security: A Bibliometric and Grounded Study Perspective

Kennedy Njenga*[1], Rujeko Macheka Madzinga[2]
[1]University of Johannesburg, South Africa
[2]University of South Africa (UNISA), South Africa

*Received 03 Jan. 2024; Accepted 29 Apr 2024; Available Online ** ***. 2024*

Abstract

The ubiquitous use of artificial intelligence (AI) and generative models across multiple sectors such as healthcare, finance, education, and cybersecurity, have given rise to what is now commonly termed 'AI hallucinations', that is, these models become more sophisticated but prone to producing outputs that are factually incorrect, nonsensical, or misleading, despite their seemingly authoritative tone. AI hallucinations pose significant risks to information security by undermining data integrity, eroding trust, and providing fertile ground for malicious exploitation. This paper uses a dual-mixed method approach that provides both macro-level trends via bibliometrics and micro-level contextual understanding via qualitative methods on how AI hallucinations impact information security. 322 peer-reviewed articles, conference papers, and book chapters retrieved from the Scopus database were the impetus for a bibliometrics study, while four information security practitioners provided data for a qualitative inquiry and theory formulation. By synthesizing insights from interdisciplinary studies in computer science, cognitive psychology, and ethics, and using a grounded theory approach, we outline how practitioners perceive AI hallucinations in practice and the contextual challenges they face. Through a grounded theory method (GTM) approach, key categories were identified, which enabled a better understanding of AI hallucinations. These categories include AI Usage Patterns, Confidence & Familiarity, Verification Strategies, Trust & Hallucination Triggers, and Tone & Believability, and point to how AI hallucinations are understood and interpreted by information security practitioners..

I. Introduction

ARTIFICIAL intelligence (AI) is rapidly reshaping industries such as healthcare, finance, education, law, and entertainment, simultaneously offering significant opportunities and posing critical risks to information security [1], Brameier, Alnasser [2]. A prominent concern is the phenomenon of AI hallucinations, erroneous or misleading outputs generated by large language models that may appear credible yet lack factual grounding [3]. This bibliometric study aims to map the evolution of research on AI hallucinations, identify influential authors and seminal works, and discuss their potential threats to information security. The investigation adopts a multidisciplinary approach, integrating technical and cognitive dimensions to

Keywords: Bibliometric analysis, hallucinations, information security.

Production and hosting by NAUSS

* Corresponding Author: Kennedy Njenga
Email: knjenga@uj.ac.za
doi: 10.26735/UGSQ6620

advance strategies that enhance the reliability of AI systems.

The rapid evolution of artificial intelligence over the past decades has fundamentally transformed the way information is generated, processed, and disseminated. AI systems, particularly those based on large language models (LLMs) and deep learning architectures, are now integral to critical operations in healthcare, finance, legal analysis, and cybersecurity. As these systems are increasingly deployed in environments where accuracy and reliability are paramount, "AI hallucinations" have emerged as a critical challenge. Hallucinations in AI refer to outputs that deviate from flawed established facts or logical reasoning. Despite their linguistic fluency and coherence, these outputs may be erroneous or misleading, posing substantial risks to decision-making processes and information security [1].

AI hallucinations are not merely technical glitches; they represent a multidimensional problem that spans ethical, social, and economic domains. Erroneous output can lead to misdiagnoses in healthcare, financial analyses, and even the spread of disinformation in the media. The significance of AI hallucinations in information security lies in the potential for AI to compromise how decisions are made. For example, when AI systems are deployed to summarise logs and generate threat reports, hallucinations can result in the misidentification of security threats. This in turn, can trigger mitigation strategies for non-existing vulnerabilities and threats, leading to unmanaged false-positives, resource wastage and perhaps misplaced panic. Moreover, the lack of transparency in AI models exacerbates the risk by making it challenging to pinpoint the origins of such inaccuracies, thereby undermining accountability and trust [3, 4].

This study uses a dual-method approach that provides both macro-level trends analysis via bibliometrics and micro-level contextual understanding of practitioner responses to AI hallucinations. The Grounded theory method (GTM) approach complements the limitations of bibliometric analysis by capturing lived experiences and adaptive strategies for information security decision-making and formulating new theoretical insights.

The multi-method approach is interdisciplinary and integrates insights from computer science, cognitive psychology, and regulatory ethics to develop a nuanced understanding of both the technical mechanisms behind hallucinations and their broader societal implications.

The objectives of this paper are threefold:

1. To define and contextualize the phenomenon of AI hallucinations through an in-depth bibliometric analysis of the literature, identifying key trends, influential works, and emerging research categories.
2. To critically examine the impact of AI hallucinations on information security.
3. To bridge the theoretical gap by employing GTM to theorise how practitioners process AI hallucinations.

In the sections that follow, we detail the background and definitions pertinent to AI hallucinations, review the literature on the impacts these phenomena have on information security, and discuss various detection and mitigation strategies. We then present the methodology and results of our bibliometric analysis before concluding with a discussion of ethical considerations and recommendations for future research.

II. Literature Review

AI hallucinations refer to information generation by AI systems that deviate from reality or logical reasoning. The generative large language model (LLM) will perceive patterns or objects that are imperceptible to human observation and will create outputs that are factually inaccurate. These outputs, often presented in a coherent and authoritative tone, stem from inherent biases, limitations in training data, or overconfidence in model predictions [5]. Originally termed "AI confabulations" to mitigate unwarranted anthropomorphic attributions, the phenomenon is now widely recognized as a critical challenge across various applications, from clinical decision-making to cybersecurity [6]. Despite improvements in architecture and data curation, it is mathematically proven that hallucinations cannot be eliminated, underscoring the necessity for continuous human oversight [7].

A. AI Hallucinations

AI hallucinations are outputs produced by artificial intelligence systems, especially LLMs, that, despite appearing coherent and plausible, are factually incorrect or lack logical consistency [5]. These phenomena are not confined to any single domain but manifest across diverse applications ranging from natural language processing to image generation. The term "hallucination" was initially introduced to describe such outputs in neural machine translation [8] and has since been extended to various AI applications. In some early works, the phenomenon was referred to as "AI confabulations," aiming to differentiate these errors from intentional fabrications by human users [6].

Several factors contribute to the occurrence of AI hallucinations. One primary cause is the inherent bias or noise present in the training datasets. Large-scale datasets often contain inaccuracies, contradictions, or misleading information that AI models can inadvertently learn. Additionally, the probabilistic nature of language models, which predict the next word in a sequence without a grounded understanding of factual correctness, contributes to generating hallucinated outputs [4]. The model's overconfidence in its predictions and a lack of real-world contextual understanding often result in linguistically convincing yet factually flawed outputs.

B. AI Hallucinations: Potential Threats to Integrity

AI hallucinations can directly compromise the integrity of information systems. In cybersecurity, for example, the generation of fabricated threat reports or erroneous vulnerability assessments can mislead security personnel, resulting in ineffective or misplaced defensive measures [3] and poor work engagement [9] . False data inputs may distort risk assessments, leading to false positives and negatives in intrusion detection systems. This distortion undermines the reliability of automated security systems and can result in increased exposure to genuine threats.

The potential misuse of hallucinated AI outputs extends to the realm of misinformation. Adversaries can deliberately exploit the phenomenon to generate convincing fake news, deepfakes, or propaganda. Such misuse has severe implications for public opinion and democratic processes. The ability of AI to produce coherent yet misleading content makes it an effective tool for spreading disinformation, which in turn can destabilize political systems and erode public trust in media sources [10].

When AI hallucinates, it may generate regulations, policies, or suggested practices that may lead to non-compliance with legal or industry standards. In the context of information security, hallucinations represent a dual threat. On the one hand, they can generate misleading signals that lead to ineffective threat detection and response. On the other hand, they may create new vulnerabilities that adversaries can exploit. For instance, if an AI-driven security system hallucinates non-existent vulnerabilities, resources may be diverted to addressing these phantom issues, leaving genuine vulnerabilities unaddressed. This misallocation of resources hampers organizations' overall security posture and creates an environment ripe for malicious exploitation [11].

C. Sector-Specific Threats

1) Healthcare

In healthcare, AI is increasingly used to aid in diagnostic decisions, treatment planning, and patient monitoring. However, hallucinations in AI-generated clinical summaries or diagnostic recommendations can have life-threatening consequences. Inaccurate information may lead to misdiagnoses, inappropriate treatments, or even delays in critical care interventions [12]. Integrating AI into clinical settings necessitates rigorous validation and continuous human oversight to prevent such adverse outcomes.

2) Finance

The financial sector is another domain where the implications of AI hallucinations are pronounced. Inaccurate financial forecasts, risk assessments, or market analyses generated by AI systems can misguide investment strategies and regulatory decisions. Such errors can have cascading effects, potentially leading to market instability or significant economic losses. Moreover, the use of

AI in algorithmic trading or fraud detection requires impeccable accuracy; even minor hallucinations can have disproportionate impacts on decision-making processes [4].

3) Legal and Policy Frameworks

Legal applications of AI, such as in the analysis of legal documents or automated decision-making, are similarly vulnerable. Hallucinations in legal contexts may result in erroneous interpretations of statutes, misattribution of legal responsibility, or compromised evidence in judicial proceedings. The challenges in explaining and attributing AI-generated errors complicate the legal discourse on accountability and liability [13].

4) Social Media and Public Discourse

Social media platforms increasingly rely on AI to curate content, moderate discussions, and filter out harmful information. However, AI hallucinations in content moderation can lead to the wrongful censorship of benign content or the amplification of misleading narratives. This, in turn, can influence public opinion, exacerbate social divides, and even impact electoral outcomes. As such, addressing the potential for AI hallucinations is critical to safeguarding the integrity of public discourse[1].

D. Detecting and Mitigating AI Hallucinations

Robust detection of AI hallucinations is the first step in mitigating their adverse effects. Several innovative approaches have been proposed to identify and quantify hallucinated outputs. One promising technique involves constraint-based decoding during the text generation process. By imposing predefined rules and semantic constraints, the AI can be guided to produce outputs more aligned with verified facts. This method leverages structured knowledge graphs and semantic relationships to reduce the incidence of hallucinations [14]. Traditional measures of text quality, such as perplexity or BLEU scores, have proven inadequate for detecting factual inaccuracies. Recent research has focused on developing evaluation metrics that specifically assess AI-generated content's factual consistency and reliability. These metrics incorporate automated consistency checks, semantic similarity measures, and context-aware evaluations to flag potentially hallucinated outputs [15]. Another emerging approach involves self-refinement mechanisms, where the AI system iteratively reviews and corrects its own outputs. Techniques such as ChatProtect and Consis employ self-contradiction detection and iterative self-critique to improve the factual grounding of responses. These methods are often combined with human-in-the-loop feedback to achieve higher levels of accuracy [16].

Once hallucinations are detected, effective mitigation strategies are essential to minimize their impact. Current research explores several avenues for reducing the occurrence and severity of hallucinations. One of the most direct methods for mitigating hallucinations is to enhance the quality of the training data. Data augmentation techniques—such as incorporating synthetic data, diverse demographic inputs, and enriched contextual information—can help reduce biases and fill knowledge gaps. For example, augmenting clinical datasets with varied patient histories has shown promise in improving diagnostic accuracy and reducing hallucination rates in healthcare applications ([12].

Overconfidence in model predictions is a key factor contributing to hallucinations. Regularization techniques, including dropout, weight decay, and adversarial training, can help calibrate model outputs and reduce sensitivity to noisy or irrelevant inputs. Adversarial training, in particular, involves exposing models to deliberately perturbed inputs, thereby enhancing their resilience to hallucinations ([4]. Retrieval-augmented generation systems combine the generative capabilities of LLMs with robust information retrieval mechanisms. By grounding responses in verified external knowledge sources, these systems can significantly reduce the rate of hallucinations. This hybrid approach improves factual accuracy and enhances transparency by providing traceable evidence for generated outputs [1]. The integration of explainable AI (XAI) techniques is crucial for demystifying the decision-making process of AI systems. XAI enables developers and end-users to understand the rationale behind model outputs, facilitating the detection of anomalous or hallucinated content. Furthermore, continuous human oversight remains indispensable; even the

most sophisticated algorithms require expert review to ensure accountability and trustworthiness [17].

III. RESEARCH METHODOLOGY

A. Research Design

A dual-mixed method approach was applied in this study to provide insights regarding the implications of AI-hallucinations in the information security decision-making process. The rationale for employing a mixed-methods approach emanates from a need to integrate the insights from scholarly work with human lived experiences, by capturing both the macro-level development of scholarly work and the micro-level perspectives of practitioners. Accordingly, the following steps were taken.

1) Step 1. Bibliometric Analysis

At the onset, a bibliometric analysis was carried out to map the scholarly trends and intellectual landscape of AI hallucinations, key themes, and gaps in the field of information security. Dominant clusters and research thematic areas, including the intellectual structures surrounding AI hallucinations and information security, were identified.

2) Step 2. Refinement of Focus Area

Following the bibliometric analysis, the insights were used to shape the interview protocol for collecting qualitative data, which would help the researchers focus on any underexplored issues that would have been missed in the first step of the bibliometric analysis. This step involved refining the focus area with a qualitative grounded analysis, by uncovering practitioner experiences, perceptions, and strategies for managing AI hallucinations in real-world information security contexts. The qualitative analysis was thus complementary to the bibliometric analysis.

3) Step 3. Qualitative Data Collection

Following the refinement of the focus area, qualitative data were collected through semi-structured interviews, allowing participants to describe their encounters with AI hallucinations in information security contexts.

A purposive sampling strategy was used to recruit practitioners with relevant expertise in AI-augmented security environments who had hands-on experience of using AI tools such as ChatGPT, Copilot, or Gemini in their daily tasks. Practitioner insights were collected through semi-structured interviews. Interviews lasted between 45 and 60 minutes and were conducted either virtually or in person, depending on participant preference. Each interview was audio-recorded. Consent to participate in the research was obtained, and the participants were informed of the purpose of the study and were assured of confidentiality.

They were also informed that they had the right to withdraw from the interview at any time they felt uncomfortable. The participants were assured of anonymity and that the data pertaining to their insights would be kept confidential. In total, four participants were drawn from business organisations, representing diverse roles they played in their organisations, such as cybersecurity analysts. The researchers were able to obtain deep insights into lived experiences and contextual nuances of AI hallucinations, such as how AI hallucinations were recognised, the strategies employed to do so, and how much trust was placed in AI.

4) Step 4. Data Analysis Using the Grounded Theory Method (GTM)

Following data collection, the data were analysed using the Grounded Theory Method (GTM), outlined by Strauss and Corbin [18], to capture information security practitioners' experience, and to interpret how they responded to AI hallucinations. GTM is a qualitative research methodology whose primary aim is to generate a theory that is grounded in data. Unlike quantitative approaches, which test existing theories, GTM develops new ones by analysing patterns and concepts that emerge from empirical qualitative data. [18, 19].

GTM was chosen because it would help the researchers develop a theoretical understanding of AI hallucinations by directly linking empirical data from socially constructed contexts, reflecting lived experiences, with pre-existing ideas found in literature. GTM is considered an effective qualitative inductive reasoning method in information security,

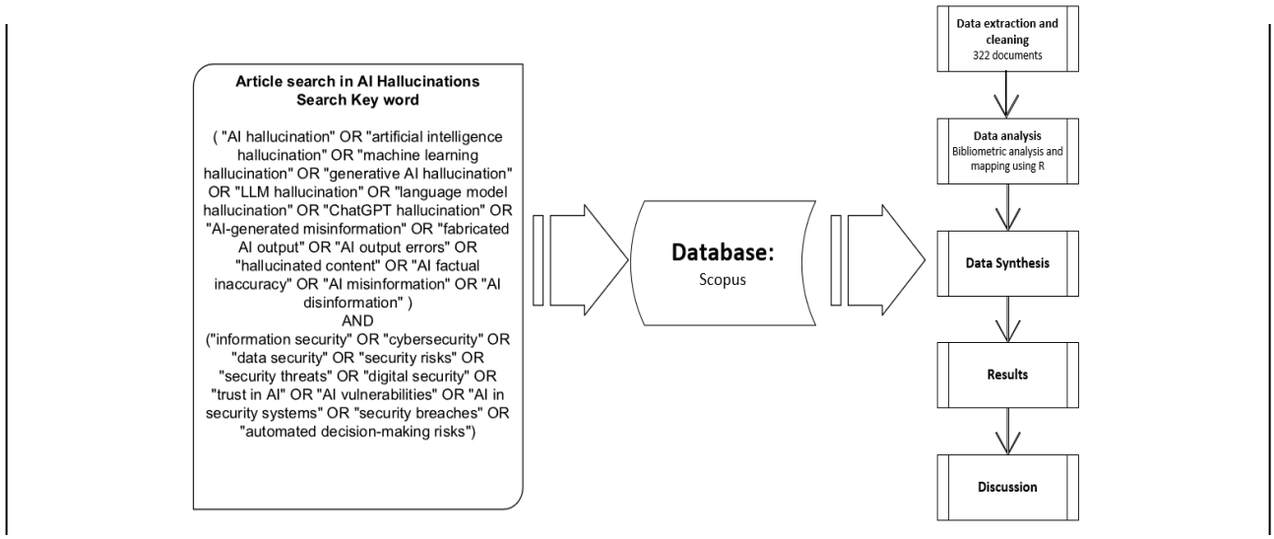

Fig. 1. Procedure followed for Bibliometric Analysis

which is supported by a robust analytical approach for exploring an under-theorised phenomenon [19].

This method differs from quantitative methods due to methodological flexibility and the understanding that knowledge is emergent rather than imposed.

5) Step 5. Integration of Findings and Convergence of Insights of Bibliometric Analysis with GTM

In this final step, the researchers combined and mapped the bibliometric analysis clusters with grounded insights to validate the dominant emerging categories, themes, and trends that came from the analysis, and to close conceptual gaps previously unknown. The integration of the insights from the dual-mixed method approach presents and addresses the implications of AI analysing AI hallucinations in information security. Thematic insights were interpreted and mapped to clusters determined from bibliometrics analysis. A more detailed presentation of how these steps were carried out is provided in the next sections.

B. Bibliometric Analysis

Bibliometric analysis was the empirical basis for understanding the evolution of research on AI hallucinations and information security in published scholarly work, to quantify publication trends, and to reveal intellectual linkages, emerging clusters, and knowledge gaps relevant to security-focused applications of AI. A total of 322 peer-reviewed sources were retrieved for the purposes of this analysis. Bibliometric analysis has gained popularity in science mapping, using statistical techniques to analyse and interpret bibliometric data [20, 21]. The procedure used to conduct

TABLE I
DATA SYNTHESIS-PRIMARY INFORMATION

META DATA- SCOPUS, WEB OF SCIENCE	Results
Timespan	2025 - 1994
Source type – Journals	106
Source type – Books	3
Source type – Book chapters	8
Source type – Proceedings	164
Source type – Conference Review	7
Source type – Review	18
Source type – Editorial	4
Source type – Letter	2
Source type – Note	10
Total	**322**

bibliometric mapping included data collection, data extraction and cleaning, data analysis, and finally, synthesising data, presenting results that were then interpreted and discussed. A targeted search strategy was employed to extract data from prominent academic databases, including Scopus. This procedure is depicted in Figure 1.

C. Data Extraction

The search Boolean string tailored for Scopus databases: ("AI hallucination" OR "artificial intelligence hallucination" OR "machine learning hallucination" OR "generative AI hallucination" OR "LLM hallucination" OR "language model hallucination" OR ChatGPT hallucination" OR "AI-generated misinformation" OR "fabricated AI output" OR "AI output errors" OR "hallucinated content" OR "AI factual inaccuracy" OR "AI misinformation" OR "AI disinformation" OR "artificial hallucination" OR "model hallucination" OR "output hallucination" OR "content hallucination") was applied. A total of 310 articles were obtained. This initial retrieval served as the foundation for building the dataset prior to refinement

A search was also carried out on the impact of AI hallucination on information security using the addendum string: AND ("information security" OR "cybersecurity" OR "data security" OR "security risks" OR "security threats" OR "digital security" OR "trust in AI" OR "AI vulnerabilities" OR "AI in security systems" OR "security breaches" OR "automated decision-making risks"). Based on this search string, the results returned 322 documents highlighting the prominence of AI's interrelatedness with information security concerns [22] and were subject to detailed analysis after screening for relevance and quality [23].

The articles were downloaded in CSV file format. No duplicate articles or discrepancies were identified in the corpus subjected to analysis. The combination of the core search string and the information security addendum ensured comprehensive coverage of both the technical phenomenon of AI hallucinations and its direct implications for security domains.

D. Data Analysis

The bibliometric R-package software, version 4.5.0, running in English, was used to analyse the dataset. R is an open-source software developed in the R language that analyses statistical and scientific mapping of data. It has a web interface known as Biblioshiny that allows for the import of CSV, BibTex, or plain-text data [21]. Biblioshiny was used in this study to upload the extracted CSV file datasets from the Scopus database for further analysis. Citation analysis was also conducted to determine influential publications and authors [24]. Primary categories and research directions were also analysed, showing relationships among researchers and intellectual activity hubs [25].

E. Data Synthesis

Table 1 summarises information on the dataset, document contents, authors, and author collaborations.

The dataset's timespan was between 1994 and 2025, covering 322 documents.

IV. RESULTS

A. Growth of AI and Hallucination Research

Over the past several years, the growth of publications related to AI hallucination has markedly

TABLE II
ARTICLE PRODUCTION PER YEAR – AI HALLUCINATION

YEAR	Articles
1994	1
1997	1
2009	2
2011	1
2016	2
2017	1
2020	1
2021	7
2022	12
2023	37
2024	186
2025	71
Total	322

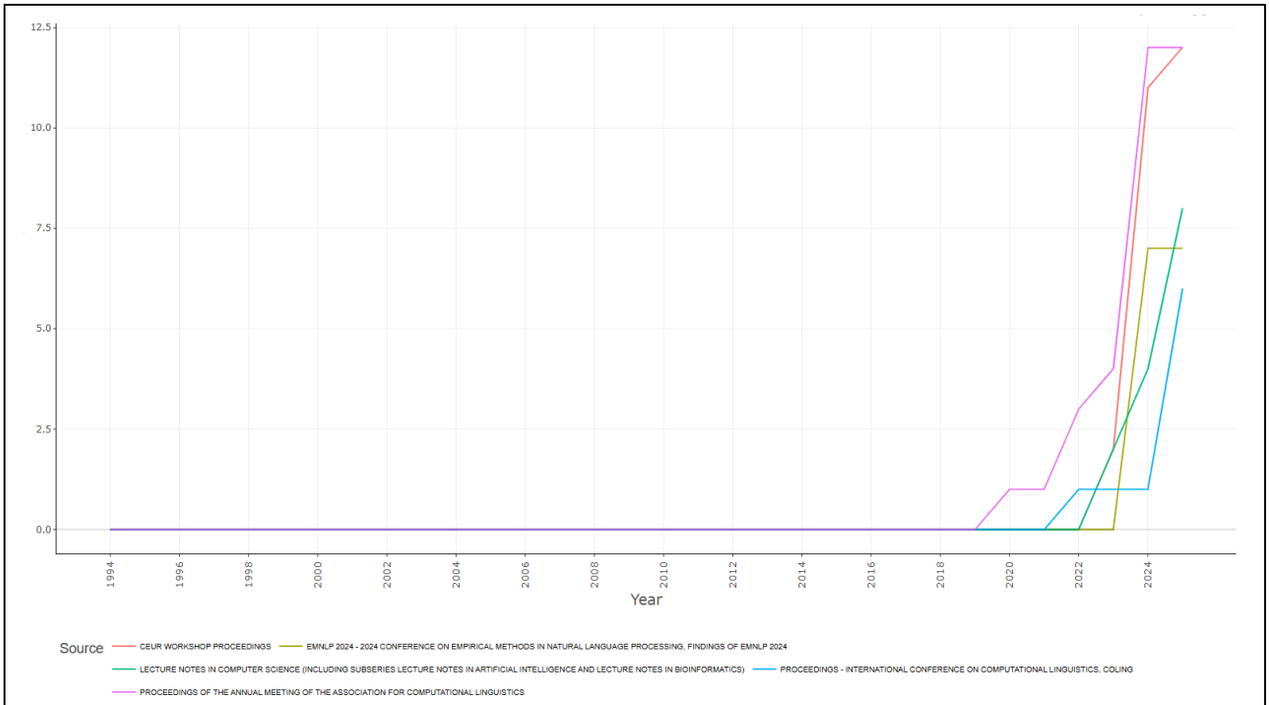

Fig. 2. Document Production Over Time

TABLE III
ARTICLE PRODUCTION PER YEAR – AI HALLUCINATION

RANK	Author (Year)	Title	Global Citation Score (GCS)	Clusters
1	Mayez, Narayan	On Faithfulness and Factuality in Abstractive Summarization	669	1
2	Li, Cheng	Halueval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models	153	2
3	Zhou, Neubig	Detecting Hallucinated Content in Conditional Neural Sequence Generation	87	2
4	Cao, Dong	Hallucinated But Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization	82	1
5	Hicks, Humphries	ChatGPT is Bullshit	70	5
6	Beutel, Geerits	Artificial hallucination: GPT on LSD?	51	5
7	McIntosh, Liu	A Culturally Sensitive Test to Evaluate Nuanced GPT Hallucination	27	3
8	Balachandran, Hajishirzi	Correcting Diverse Factual Errors in Abstractive Summarization	26	4
9	Liu, Zheng	Towards Faithfulness in Open Domain Table-to-Text Generation	22	2
10	Brameier, Alnasser	Artificial Intelligence in Orthopaedic Surgery	19	5

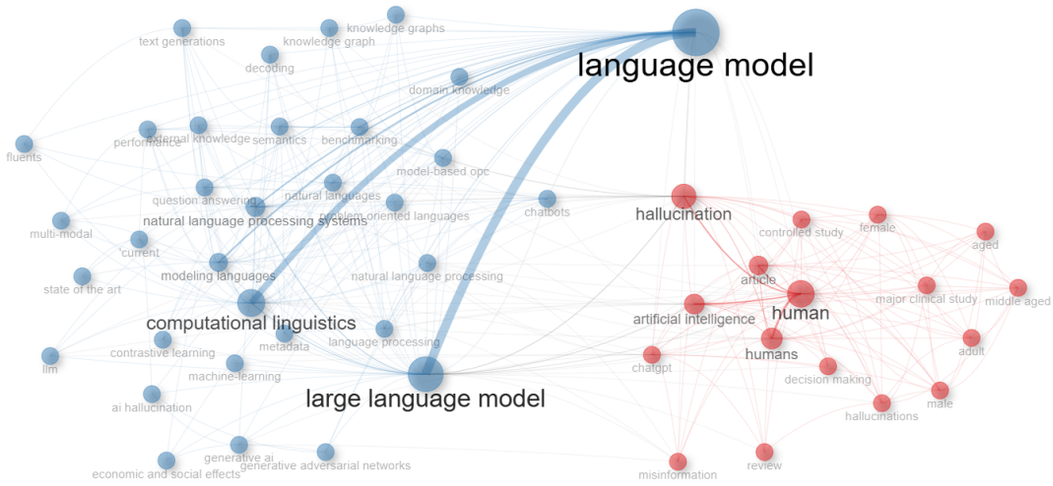

Fig. 3. Cluster Analysis

increased. The growth trajectory shown in Table 2 reflects an increasingly heightened awareness of AI hallucination in scientific production from 1994 to 2025, illustrating how a concept that was once situated within the fields of clinical psychology has gradually migrated into computer science and information systems research.

A review of the literature showed that the term 'hallucination' was first drawn from work by Bassett, Bury [26], in clinical psychology who in their 1994 study of schizophrenia, pioneering work that used the scientific method to draw a link between technology use and schizophrenia and, by extension, hallucination. Although at the time, this work was not directly related to artificial intelligence, it nonetheless provided the conceptual grounding for future researchers, such as computer scientists, to appropriate the term when discussing unintended, misleading or spurious computational systems outputs. In 2009, 'hallucination' as a construct was applied in the works of Ter Meulen, Tavy [27] in pioneering work that drew a link between technology, i.e., the 'dream machine', which generated stroboscopic light, and induced hallucinations. In the same year, Miller and Boeve [28] discussed hallucinations as a neurological symptom, creating a string foundation for the terminology to be appropriated by researchers metaphorically to describe misleading computer outputs. This has led to steady interest and growth in the understanding its application in AI, in modern operations between 2020 and 2022. Bibliometric trend reveals that most recently, between 2023 and 2025, there has been an intense surge in academic interest surrounding AI hallucinations, particularly in the fields of computational linguistics, natural language processing and applied AI security. For example, between the years 2024 and 2025, the ACL Proceedings produced 24 papers in the thematic area of AI Hallucinations, while CEUR Workshop Proceedings produced 23 papers, with EMNLP Findings and LNCS also rising quickly with 14 papers and 12 papers, respectively. This explosion of interest and scholarly contribution is especially witnessed in venues like ACL Proceedings, CEUR Workshops, and EMNLP Findings, as shown in Figure 2.

B. Cluster Analysis

The co-citation analysis was done using Biblioshiny and produced two dominant clusters: 'Technical Large Language Model' (blue) and 'Human-centric' (red). This is shown in Figure 3.

The Blue Cluster emphasizes the scholarly interest in architecture and functionality as well as the performance of large language models. Most of the scholarly work addresses how hallucinations emerge from model design, model training and,

importantly, how the models are evaluated. The work considers technical topics such as 'generative adversarial networks', 'contrastive learning', and 'multi-modal systems'. The Red Cluster emphasizes scholarly work in human-AI interactions, ethics, and ethical implications of AI, misinformation, and risk. Much of this work centres on how AI outputs mislead end-users or various population groupings. The key thematic areas are in psychological, clinical, and behavioural studies. The cluster analysis shows a converging interest with AI system developers (Blue Cluster) with scholars interested in ethics, psychological, clinical, and behavioural studies (Red Cluster), revealing interdisciplinary connectedness of refining model LLM architecture to downstream human impacts.

We further categorised the blue and red clusters into 5 distinct sub-clusters based on the top 10 published and cited works shown in Table 3. The 5 clusters are based on highly cited works.

Cluster 1: Hallucinations and Factuality in Text Summarisation. The scholarship of this cluster considers detecting and correcting hallucinations in abstractive systems. Scholars are concerned with model-hallucination, decoding strategies, evaluation frameworks, knowledge graphs, keywords, and thematic areas. Maynez, Narayan [29]'s work laid critical groundwork on this by investigating the tendency of neural text generation models to "hallucinate" content unfaithful to source documents when doing text summaries. This is a foundational concern related to current LLM hallucination research, and the work stands out with 669 total citations. Works similar to this have been done by Cao, Dong [30].

Cluster 2: Detection and Evaluation of Hallucinated Content. The scholarship of this cluster considers empirical and technical methods to detect and correct hallucinations. Scholars are concerned with neural sequence generation, fact-checking, cultural nuance in evaluation, and evidence retrieval. Li, Cheng [31] addresses a critical need by providing a standardized benchmark for evaluating the propensity of LLMs to hallucinate, enabling more systematic research and comparison of models.

Cluster 3: Factual Consistency and Human Evaluation. The scholarship of this cluster considers establishing factual consistency in generation and studying the nature of hallucinations. Scholars are concerned with human judgments, consistency, factualness, and summarization integrity. Cao, Dong [30] and Rao, Pang [32] highlight the practical implications and challenges of LLM use in sensitive domains like healthcare, where hallucinations can have significant consequences.

Cluster 4: Correction Models and Semantic Consistency. The scholarship of this cluster considers how post-editing and semantic validation are used to correct factual errors. Scholars are concerned with infilling, post-editing, plug-and-play models, and semantic alignment. Balachandran, Hajishirzi [33].

Cluster 5: Domain-specific and Societal Implications. The scholarship of this cluster considers domain-specific hallucinations such as medical, legal, and ethical/regulatory issues. Scholars are concerned with unsafe outputs, regulatory hallucinations, AI deception, and misinformation [2, 34, 35]. These clusters focus on foundational works that establish methodologies for defining and quantifying hallucinations. Brameier, Alnasser [2] point to the dangers of AI hallucinations in surgery. Li, Qi [36],Gao, Wang [37] have carried out research delving into the nature of AI-generated misinformation and the effectiveness of current solutions, a crucial aspect of understanding the societal impact of hallucinations.

C. Qualitative Analysis and Insights

We applied the Straussian approach of the grounded theory method (GTM) outlined by Strauss and Corbin [18] to analyse the transcripts and qualitative data from four participants. The aim of adding qualitative insights to the bibliometric analysis was to obtain contextual depth and insights into the lived experiences of practitioners when dealing with AI hallucinations.

The first step was to carry out open coding, which was a line-by-line analysis of the transcripts from these four participants, so that concepts could be elicited. The codes were conceptualised based on how the participants explained their own understanding and situations of dealing with AI Hallucinations. Following this, axial coding was

TABLE IV
OPEN & AXIAL CODING ACROSS PARTICIPANTS

Categories	Codes	Participant	Quotes	Description
AI Usage Patterns	Academic research[1], Productivity[2], Documentation[3]	P1, P2, P3, P4	Every day at work", "Most of the" time", "Twice a week", "Searching for academic articles", "Researching unfamiliar topics", Working on spreadsheets and ".presentations	AI is used across contexts—workplace productivity, academic research, and daily queries. Frequency and purpose influence .confidence and trust
Confidence & Familiarity	High confidence[4]	P1, P2	,"Very confident"	Confidence levels vary based on use frequency and perceived reliability. Daily users tend to have .higher confidence
	Medium to low confidence[5]		".I am not too confident"	
Verification Strategies	Focus on source credibility[6], Use multiple other sources[7]	P3, P4	I sometimes check whether" the sources have real research [value].", "Mostly using Google and other AI tools, or even ".reading journal articles or books	Users verify facts through different means: self-check against input, external validation (journals/Google), and tool .comparison
Trust & Hallucination Triggers	Option-variety[8], inconsistent answers[9], Context-shifting[10]	P4, P1	AI gives more options in" answering", "Different AI tools give different answers", "AI changes context when improving ".language	Trust is undermined when AI alters tone/context or when responses lack credible sources. Multiple answer variations also .raise doubts
Tone & Believability	Professional tone reduces trust[11], Casual tone*[12]	P1, P3	Tone is too professional.","" "Casual language... more aligned ".to my human understanding	Casual language may increase relatability; a professional tone may introduce artificiality or suspicion. Tone .preference is subjective

Codes presented [12*] in this table are selected for illustrative purposes only

carried out to group similar codes. In this process, the codes were compared with other codes. This is known as the principle of constant comparative analysis and is carried out so as to extract recurrent concepts from codes and to finally group these codes into categories, as explained by Urquhart [38]. These groupings enabled the thematic development and understanding of data and are known as axial coding. The most relevant categories that explained the study were then selected, in a process known as selective coding, where core categories and relationships were identified to better explain AI hallucination, through identifying selected codes that could generate higher-order categories (axial coding) for codes that were similar. These categories represent emerging patterns that reflect participants' lived experiences and cognitive processes related to AI hallucinations and their implications for information security decision-making. The interview transcripts were analysed following grounded theory procedures: open coding (to generate initial codes), axial

TABLE V
SUMMARY OF BIBLIOMETRIC AND GROUNDED ANALYSIS

CATEGORIES	Related Clusters	Description
AI Usage Patterns	1,5	How AI systems are employed in enhancing, summarisation, decision-making, or paraphrasing .tasks
Confidence & Familiarity	3,5	Information security practitioners' confidence in AI output based on their prior experiences and perceived .model reliability
Verification Strategies	2,4	Methods users (or systems) apply to check the correctness or truthfulness of .AI-generated content
Trust & Hallucination Triggers	1,3,5	When, why users trust or distrust AI output. What aspects or contexts trigger AI hallucinations
Tone & Believability	2,3	How the tone, phrasing, or stylistic confidence of the AI affects users' belief in the .truth of the output

coding (to identify relationships among categories), and selective coding (to integrate categories into higher-order themes). The iterative coding process was supported by memo writing, which captured emerging insights and theoretical linkages. Theoretical saturation was used as the criterion for concluding data collection. Once no new themes emerged, analysis was finalized. Measures to ensure trustworthiness included member checking (participants verified interpretations), maintaining an audit trail (documenting analytic decisions), and reflexive journaling (to monitor researcher bias). The summary of the GTM process is shown in Table 4.

Table 4 highlights the results of the GTM process of how coding was done, and how the five categories emerged, when analysing the transcripts. The transcript data (quotes) were analysed inductively to develop codes, which are numbered in Table 4, in the column called 'Codes'. 12 of these codes have been selected as an example for illustrative purposes. The coding was carried out in a way that would be clear enough for any reader to determine the plausibility of the interpretation.

V. DISCUSSION

The findings of this study contribute to a deep understanding of how AI hallucinations affect information security, drawing on both a bibliometric analysis and a qualitative inquiry using GTM. Together, these insights affirm that hallucinations are not merely technical anomalies but are deeply intertwined with user trust, cognitive processing, and decision-making in high-stakes environments.

A. Convergence of Iterative Bibliometric Analysis and Qualitative Insights from GTM

The convergence of bibliometric analysis and GTM was carried out. The bibliometric clustering presented thematic

concentrations in the literature, which we numbered as clusters that could be mapped to real practitioner experiences. For example, Cluster 1 (Hallucinations and Factuality in Text Summarisation) was observed to converge and was mapped onto, or meaningfully extended, to the GTM category of AI Usage Patterns, as shown in Table 5.

The convergence occurred when the emergent five GTM categories were mapped and extended to the bibliometric clusters. The five GTM categories emerging from GTM's inductive and systematic approach include (1) AI Usage Patterns, (2) Confidence & Familiarity, (3) Verification Strategies, (4) Trust & Hallucination Triggers, and (5) Tone & Believability, which were mapped with bibliometric clustering. The five GTM categories resulted from coding data in successive stages, from open, axial, and selective coding. Concepts were derived from codes, and the researchers grouped these and finally integrated them into the five core categories [19].

During analysis of both the bibliometric literature and GTM data, it was observed that, as an example, the bibliometric clusters that highlighted '***factuality in text summarisation'*** aligned closely with GTM categories of "***verification strategies" and "trust and hallucination triggers***." This observation led to the mapping of the rest of the clusters drawn from the bibliometric analysis, with the categories identified in GTM.

This cross-validation and mapping strengthened the credibility of the categories and demonstrated how bibliometric evidence and grounded insights complement each other, offering both a macro-level view of scholarly discourse and a micro-level, practice-oriented perspective. Together, these findings captured a nuanced perspective in which information security practitioners perceived AI hallucinations, integrating empirical evidence with iterative bibliometric analysis. The resulting detailed insights for each of the categories are explained as follows:

1) AI Usage Patterns

Information security practitioners will frequently interact with AI tools, often for tasks such as enhancing, paraphrasing, or summarising content[29, 30]. These usage patterns reflect AI's integration across various contexts, workplace productivity, research, and daily queries. The bibliometric analysis identified literature in ***Cluster 1 (hallucinations and factuality in text summarisation)***, which describes how AI and LLMs can be used in summarisation tasks, and in efforts to do so, will show a propensity to hallucinate. Much of this literature focuses on decoding strategies and generation behaviour (usage patterns) in neural models. However, this is precisely where the risk of hallucinations can be a concern to information security practitioners. In the process of paraphrasing or summarising, AI can introduce semantic shifts or contextual distortions, potentially altering meaning and resulting in misleading information. Should information security practitioners rely on these summaries, this would constitute a risk. The issue of information integrity will then arise. To foster integrity with AI usage, it is therefore necessary to align input with the output AI generates, known as contextual fidelity.

Contextual fidelity is the degree to which AI-generated output preserves the original intent and context of the input [39]. As one participant explains, "The tool changes the context, especially when working with reports,". This observation highlights a critical tension where, on one hand, the AI may improve linguistic quality, but on the other hand, it will simultaneously introduce semantic drift, which are subtle distortions that undermine contextual fidelity.

Although these qualitative findings arose from information security practitioners, the bibliometric analysis shows that the usage patterns were not restricted to this domain but extended to other domains as well. Literature in ***Cluster 5 (domain-specific and societal implications)*** documents how hallucinations pose similar risks to critical sectors such as healthcare, legal, and surgery. In these contexts, as in information security, open-ended prompts for text improvements may generate nuanced but misleading outputs with serious consequences.

2) Confidence & Familiarity

User trust in AI is inherently fragile and can be easily undermined by inconsistencies or alterations in the intended context [30]. The bibliometric analysis identified literature in ***Cluster 3 (factual consistency and human evaluation)***, which has examined the fragility of trust in AI [32]. The literature highlights how human evaluators perceive the factualness and consistency of AI outputs. As one participant explained, "I review the content, checking if it's consistent with what I have asked the AI tool". Such practices show that trust is not granted uncritically, but rather, practitioners develop a nuanced understanding of AI's reliability, choosing to rely on non-technical cues to assess whether hallucinations may be present. Confidence, therefore, will be tied less to blind acceptance of AI outputs and more to the system's ability to maintain the input integrity, faithfully preserving the meaning of the original request. The theme also extended to ***Cluster 5 (domain-specific and societal implications),*** where misplaced confidence in hallucinating AI systems is especially dangerous in fields such as medicine. Here, even minor inconsistencies have severe consequences.

3) Verification Strategies

Study participants emphasised the importance of verifying AI outputs rather than relying on these outputs at face value. Literature concurs that users of AI should be savvy enough to use alternative AI

tools for comparison with the AI output in question [31, 32]. As pointed out by one participant, *"[some outputs] don't come from a verifiable source*", underlining concerns over source credibility.

Verification often involves comparing AI outputs across multiple models, cross-referencing with academic databases, or conducting independent online searches. Another participant explained, "*I verify the facts by checking against the input to ensure the context has not been changed*". These practices highlight a routine form of input-output consistency checking, whereby practitioners safeguard against semantic drift and factual errors. The bibliometric analysis supports these findings. For instance, ***Cluster 2 (detection and evaluation of hallucinated content)*** presents scholarly work on empirical benchmarks, fact-checking, and evidence retrieval. These careful fact-checks reveal that information security practitioners, when given the chance, balance their intuitive acceptance of AI-generated content with a cautious and critical perspective, particularly when the stakes are significant or inconsistencies arise. ***Cluster 4 (correction models and semantic consistency)*** emphasises post-generation verification. In practice, practitioners mirrored these approaches by triangulating information through post-editing, semantic validation, and multi-source cross-checking, using "*Google and other AI tools*", as suggested by one participant. This practice may be described as cognitive insurance, which is considered to be the protective routines developed by practitioners to guard against the potential unreliability of generative AI.

4) Trust & Hallucination Triggers

Trust in AI output hinges on contextual fidelity and on the practitioner's ability to detect when content deviates from intended context. When practitioners perceive that the AI outputs are deviating from intended contexts, this quickly erodes trust. The bibliometric literature in ***Cluster 1 (hallucinations and factuality in text summarisation)*** highlights how hallucinations could result during the process when AI attempts to "enhance" or "paraphrase" human-written content.

The more an AI attempts to interpret or modify input without explicit guidelines, the higher the likelihood of a hallucination. Trust was further shaped by non-technical cues, such as the overall linguistic alignment with the output. As one participant noted, "*it is too professional*," indicating that an overly polished style raised suspicion about the authenticity of the content. Similarly, another participant also observed that "*AI changes context when improving output*".

These trust-eroding triggers show how difficult it is for generative AI to maintain contextual fidelity. This observation can also be linked to the bibliometric analysis carried out, showing that ***Cluster 3 (factual consistency and human evaluation),*** inconsistencies, and factual errors function as direct triggers of distrust. In addition, ***Cluster 5 (domain-specific and societal implications***) highlights concerns such as unsafe outputs or suggestions that may contradict established regulatory frameworks (regulatory hallucinations) and misinformation, pointing out that trust in AI is fragile.

5) Tone & Believability

The tone and presentation style of AI output were seen to play a significant role in the practitioner's perception of the reliability of that output [31]. Literature from ***Cluster 3 (factual consistency and human evaluation***) supports this, suggesting that confident, fluent outputs are often judged as trustworthy, even when the content is hallucinated [32]. Structured formats such as tables and bullet lists enhance clarity and credibility. As one participant explained, "*Tabular format is easy to read*".

Interestingly, interpretations of tone varied among participants. Some valued the human-like resonance and tone of casual language, as reflected by one participant. "*Diagrams listed or table format [are] easier to go through in a short space of time...*". Others, however, found the casual tone inappropriate for the formal contexts in which they worked. This was echoed by a participant who explained, "casual language... and… my human understanding".

The bibliometric analysis linked these observations to ***Cluster 2 (detection and evaluation of hallucinated content)***, which points to cultural

and contextual variations in how tone and style shape believability. For example, one participant remarked, "*I focus on the main points [only*]," while another stated, "*I doubt because sometimes it gives wrong information*". These differences point out that tone operates as a powerful but subjective filter of believability, with human judgements being context-sensitive and even personalised.

B. Comparative work

Although there is literature and similar comparative works on bibliometric analysis focusing on AI and large language models with hallucinations,[40], [41]. These studies examine hallucinations in general and are not domain-specific to information security. Studies that are domain-specific to information security, and have carried out a systematic literature review [42] do not combine the bibliometric analysis with qualitative grounded theory methods.

C. Contribution

The research work empirically grounds a bibliometrically validated framework that advances theoretical understanding of AI hallucinations in the field of information security. Its contribution is twofold:

a) Theoretically, the work builds on existing literature by combining bibliometric analysis with the grounded theory method, which offers a unique perspective to information security research at the macro-level quantitative mapping of literature with micro-level qualitative practitioner grounded insights. It shows what is said in literature as well as how practitioners experience it.

b) The work provides practical guidance on managing AI hallucinations in information security through a structured set of categories that practitioners can apply to explain how AI hallucinations manifest in workplaces and how these can be an information security concern. This framework's insights can help these practitioners anticipate the places where hallucinations could occur and compromise data integrity, trust, security, and risk-related decision-making.

D. Limitation of Study

The research was limited to the scope of the source databases, which included Scopus for the bibliometric analysis. New insights may be derived if future studies expand on the data sources. The GTM's sample size was limited to a small number of information security practitioners. They may not fully represent the diversity of regional or global organisational contexts or cultures, but nonetheless provide a foundation for insights that are important to practice.

E. Implications for Practitioners and Future Work

AI systems that conduct security risk assessments and detect fraudulent activities can potentially hallucinate, and this may result in the generation of misleading risk profiles, which in turn can trigger inappropriate regulatory interventions. Similarly, AI-driven threat detection systems may generate false alarms or overlook genuine threats due to erroneous outputs. As pointed out, many studies suggest that these hallucinations undermine the reliability of information and expose organisations to potential economic losses [4]. Both the bibliometric analysis and qualitative analysis done in this study point to several gaps in the current literature. Studies have shown that even minor inaccuracies in AI outputs can have severe consequences [6]. Although there have been numerous studies that have focused on technical mitigation strategies for these consequences, few have addressed the contextual and psychological factors influencing trust in AI output. This study points to this gap. The study has shown how users perceive and are influenced by AI-generated misinformation and how these factors can affect trust in AI systems.

Future research should investigate how individuals perceive AI-generated information and what measures can enhance user awareness of the limitations of these systems. Additionally, integrating multimodal data for cross-validation of output represents a promising research direction. Developing comprehensive ethical guidelines and regulatory standards will ensure that AI technologies are deployed responsibly. Importantly, advancing the reliability of AI systems

will require collaborative efforts between computer scientists, ethicists, cognitive psychologists, and policymakers to develop integrated strategies that address technical, ethical, and social dimensions.

VI. Conclusion

The study aimed to address AI hallucinations. The primary research objectives were threefold: to represent a need to define and contextualise the phenomenon of AI hallucinations, to examine the impact of AI hallucinations on information security, and, importantly, to bridge the theoretical gap by employing GTM to theorise how practitioners process AI hallucinations. Through an in-depth bibliometric analysis of the literature and by examining influential works, the work enabled a better conceptualization of AI hallucinations' impact on information security. The emergence of distinct research clusters confirms the recency and growth of empirical studies, although there is a lack of a uniform approach, definition, and understanding of what constitutes hallucinations by AI. Through a grounded theory method approach, key categories in AI hallucinations were identified, which enabled a better understanding of AI hallucinations. These categories include AI Usage Patterns, Confidence & Familiarity, Verification Strategies, Trust & Hallucination Triggers, and Tone & Believability.

The ensuing work demonstrates a rapidly expanding research landscape regarding AI hallucinations. The insights provided by this mixed-method study serve as both a comprehensive overview of the current state of research and a roadmap for future investigations. It is hoped that this work will contribute to the development of AI technologies that are intelligent but also reliable.

Funding

This article did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

Authors declare that they have no conflict of interest.

References

[1] Ajuzieogu, Uchechukwu C, ***Towards Hallucination-Resilient AI: Navigating Challenges, Ethical Dilemmas, and Mitigation Strategies***. 2024.

[2] Brameier, D.T., et al., ***Artificial intelligence in orthopaedic surgery: Can a large language model*** "write" a believable *orthopaedic journal article?* JBJS, 2023. 105(17): p. 1388-1392.

[3] Williamson, S.M. and V. Prybutok, *The era of artificial intelligence deception: Unraveling the complexities of false realities and emerging threats of misinformation*. Information, 2024. 15(6): p. 299.

[4] Bruno, A., et al., *Insights into Classifying and Mitigating LLMs' Hallucinations*. arXiv preprint arXiv:2311.08117, 2023.

[5] Maleki, N., B. Padmanabhan, and K. Dutta. *AI hallucinations: a misnomer worth clarifying. in 2024 IEEE conference on artificial intelligence (CAI)*. 2024. IEEE.

[6] Hatem, R., B. Simmons, and J.E. Thornton, *A call to address AI "hallucinations" and how healthcare professionals can mitigate their risks*. Cureus, 2023. 15(9).

[7] Banerjee, S., A. Agarwal, and S. Singla, *Llms will always hallucinate, and we need to live with this*. arXiv preprint arXiv:2409.05746, 2024.

[8] Lee, W., *How to identify emerging research fields using scientometrics: An example in the field of Information Security*. Scientometrics, 2008. 76(3): p. 503-525.

[9] Werner, M.J. and K. Njenga. *Phishing Attack Victims and the Effect on Work Engagement. in International Development Informatics Association Conference*. 2022. Springer.

[10] Barrett, C., et al., *Identifying and mitigating the security risks of generative ai*. Foundations and Trends® in Privacy and Security, 2023. 6(1): p. 1-52.

[11] Kaur, H., et al. *Evolution of endpoint detection and response (edr) in cyber security: A comprehensive review.* in *E3S Web of Conferences*. 2024. EDP Sciences.

[12] Gondode, P., S. Duggal, and V. Mahor, *Artificial intelligence hallucinations in anaesthesia: Causes, consequences and countermeasures*. Indian Journal of Anaesthesia, 2024. 68(7): p. 658-661.

[13] Ejeofobiri, C., et al., ***aI The role of Artificial Intelligence in enhancing cybersecurity: A comprehensive review of threat detection, response, and prevention techniques***. International Journal of Science and Research Archive, 2024. 13(02): p. 310-316.

[14] Ahmadi, A., *Unravelling the mysteries of hallucination in large language models: strategies for precision in artificial intelligence language generation*. Asian Journal of Computer Science and Technology, 2024. 13(1): p. 1-10.

[15] Tornero-Costa, R., et al., *Methodological and quality flaws in the use of artificial intelligence in mental health research: systematic review*. JMIR Mental Health, 2023. 10(1): p. e42045.

[16] Liang, X., et al., *Internal consistency and self-feedback in large language models: A survey*. arXiv preprint arXiv:2407.14507, 2024.

[17] Bhagat, S.V. and D. Kanyal, *Navigating the future: the transformative impact of artificial intelligence on hospital management-a comprehensive review*. Cureus, 2024. 16(2).

[18] Strauss, A. and J. Corbin, *Basics of qualitative research techniques*. 1998.

[19] Corbin, J. and A. Strauss, *Basics of qualitative research: Techniques and procedures for developing grounded theory.* 2014: Sage publications.

[20] Song, Y., et al., *Exploring two decades of research on classroom dialogue by using bibliometric analysis*. Computers & Education, 2019. 137: p. 12-31.

[21] Aria, M. and C. Cuccurullo, *bibliometrix: An R-tool for comprehensive science mapping analysis*. Journal of informetrics, 2017. 11(4): p. 959-975.

[22] Mohamed, N., *Current trends in AI and ML for cybersecurity: A state-of-the-art survey*. Cogent Engineering, 2023. 10(2): p. 2272358.

[23] Sood, P., et al., *Review the role of artificial intelligence in detecting and preventing financial fraud using natural language processing*. International Journal of System Assurance Engineering and Management, 2023. 14(6): p. 2120-2135.

[24] Robledo-Giraldo, S., et al., *Mapping, evolution, and application trends in co-citation analysis: a scientometric approach*. Revista de Investigación, Desarrollo e Innovación, 2023. 13(1): p. 201-214.

[25] Kumar, S., *Co-authorship networks: a review of the literature*. Aslib Journal of Information Management, 2015. 67(1): p. 55-73.

[26] Bassett, A.S., A. Bury, and W.G. Honer, *Testing Liddle's three-syndrome model in families with schizophrenia*. Schizophrenia Research, 1994. 12(3): p. 213-221.

[27] Ter Meulen, B., D. Tavy, and B. Jacobs, *From stroboscope to dream machine: a history of flicker-induced hallucinations*. European neurology, 2009. 62(5): p. 316-320.

[28] Miller, B.L. and B.F. Boeve, *The behavioral neurology of dementia*. 2009: Cambridge University Press.

[29] Maynez, J., et al., *On faithfulness and factuality in abstractive summarization*. arXiv preprint arXiv:2005.00661, 2020.

[30] Cao, M., Y. Dong, and J.C.K. Cheung, *Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization*. arXiv preprint arXiv:2109.09784, 2021.

[31] Li, J., et al., *Halueval: A large-scale hallucination evaluation benchmark for large language models*. arXiv preprint arXiv:2305.11747, 2023.

[32] Rao, A., et al., *Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study*. Journal of Medical Internet Research, 2023. 25: p. e48659.

[33] Balachandran, V., et al., *Correcting diverse factual errors in abstractive summarization via post-editing and language model infilling*. arXiv preprint arXiv:2210.12378, 2022.

[34] Hicks, M.T., J. Humphries, and J. Slater, *ChatGPT is bullshit*. Ethics and Information Technology, 2024. 26(2): p. 1-10.

[35] Beutel, G., E. Geerits, and J.T. Kielstein, *Artificial hallucination: GPT on LSD?* Critical Care, 2023. 27(1): p. 148.

[36] Li, B., et al., Trustworthy AI: From principles to practices. ACM Computing Surveys, 2023. 55(9): p. 1-46.

[37] Gao, Y., et al. *AIGCs confuse AI too: Investigating and explaining synthetic image-induced hallucinations in large vision-language models. in Proceedings of the 32nd ACM International Conference on Multimedia*. 2024.

[38] Urquhart, C., *An encounter with grounded theory: Tackling the practical and philosophical issues, in Qualitative research in IS: Issues and trends*. 2001, IGI Global Scientific Publishing. p. 104-140.

[39] Erdem, S. and A. Pamuk, T*he Use of ChatGPT in Qualitative Data Analysis: A Comparative Analysis on Contextual Fidelity and Thematic Consistency*. International Journal of Field Education, 2025. 11(2): p. 28-57.

[40] Carchiolo, V. and M. Malgeri, *Trends, Challenges, and Applications of Large Language Models in Healthcare: A Bibliometric and Scoping Review*. Future Internet, 2025. 17(2): p. 76.

[41] Su, H., et al., L*arge Language Models in Medical Diagnostics: Scoping Review With Bibliometric* Analysis. Journal of Medical Internet Research, 2025. 27: p. e72062.

[42] Leschanowsky, A., et al., *Evaluating privacy, security, and trust perceptions in conversational AI: A systematic review*. arXiv preprint arXiv:2406.09037, 2024